Scrapy与Redis的结合使用

# scrapy+redis实现分布式爬虫

***scrapy***是Python界出名的一个爬虫框架。Scrapy是一个为了爬取网站数据，提取结构性数据而编写的应用框架。 可以应用在包括数据挖掘，信息处理或存储历史数据等一系列的程序中。

虽然scrapy能做的事情很多，但是要做到大规模的分布式应用则捉襟见肘。有能人改变了scrapy的队列调度，将起始的网址从start\_urls里分离出来，改为从Redis读取，多个客户端可以同时读取同一个redis，从而**实现了分布式的爬虫**。就算在同一台电脑上，也可以多进程的运行爬虫，在大规模抓取的过程中非常有效。

# 学习步骤介绍



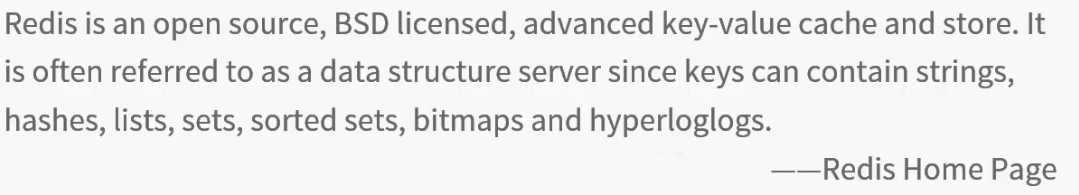
# Redis介绍

## Redis的logo



## Redis简单介绍



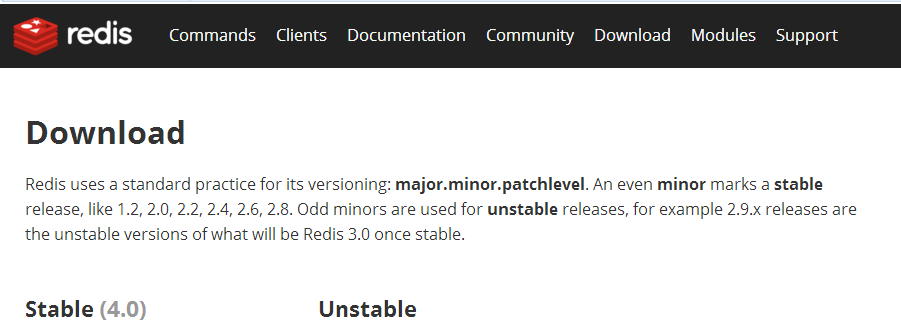


## Redis的安装

### 下载windows的redis安装包

Redis的官网地址：<https://redis.io/>

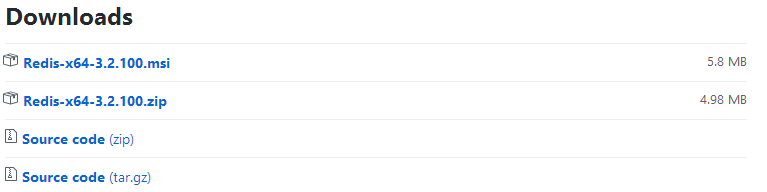
点击**Download选项**：



但是，Redis官方默认的是Linux安装包，对于Windows安装包官方没有安装包。

mis的安装包在GitHub上面有：

<https://github.com/MicrosoftArchive/redis/releases>



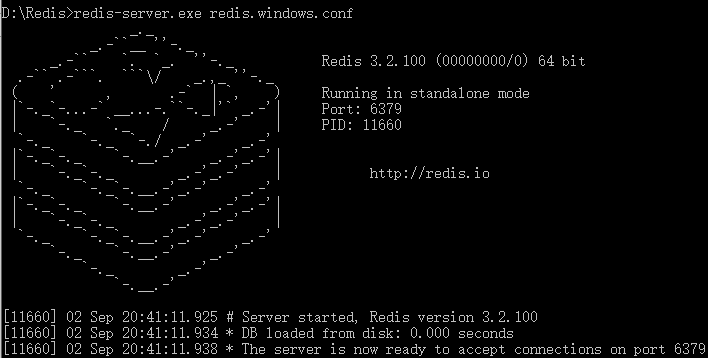
<https://github.com/MicrosoftArchive/redis/releases/download/win-3.2.100/Redis-x64-3.2.100.msi>

### 配置环境变量

安装后，配置环境变量，将redis的目录添加到Path路径下。

### 启动redis-server

打开命令行（**Win+R**），进入redis的目录**cd D://Redis/**，然后输入命令**redis-server.exe redis.windows.conf** 即可启动。



小技巧：

在D://Redis/目录下，创建startup.bat文件，文件内容为

**redis-server redis.windows.conf**

双击执行该文件，就可以打开**redis-server**。

### 启动redis-cli：

启动redis-server之后，再启动一个命令行Win+R，输入**redis-cli**即可进入**redis的客户端**，输入exit即可退出redis客户端。

**清空缓存：redis-cli flushdb**

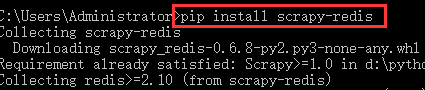
**也可以在D://Redis/目录下，创建clear.bat文件，文件内容为**

**redis-cli flushdb**

**双击执行该文件，就可以清空redis。**

## Python中scrapy-redis模块的安装

### 方法1：pip install scrapy-redis

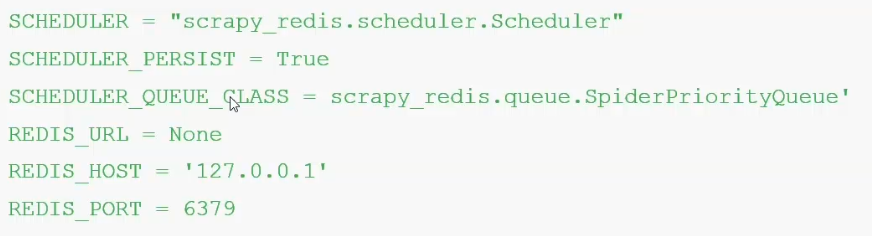


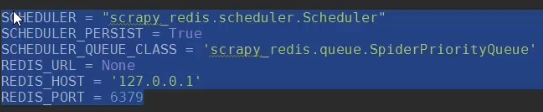


### 方法2：下载whl安装包

**pip install \*\*\*.whl**

## 在scrappy的setting.py文件中配置redis





# 项目实战

参考：

<http://blog.csdn.net/howtogetout/article/details/51633814>

